

Pentaho Data Integration

3 j (21 heures)

Ref : PDI

Public

Toute personne amenée à gérer des flux de données inter-application et / ou ayant en charge l'alimentation d'une base de données décisionnelle

Pré-requis

Toute personne amenée à gérer des flux de données inter-application et / ou ayant en charge l'alimentation d'une base de données décisionnelle

Moyens pédagogiques

Formation réalisée en présentiel ou à distance selon la formule retenue
Exposés, cas pratiques, synthèse, assistance post-formation pendant trois mois
Un poste par stagiaire, vidéoprojecteur, support de cours fourni à chaque stagiaire

Modalités de suivi et d'évaluation

Feuille de présence émargée par demi-journée par les stagiaires et le formateur
Exercices de mise en pratique ou quiz de connaissances tout au long de la formation permettant de mesurer la progression des stagiaires
Questionnaire d'évaluation de la satisfaction en fin de stage
Auto-évaluation des acquis de la formation par les stagiaires
Attestation de fin de formation

Cette formation Pentaho Data Integration permettra aux participants d'appréhender la notion d'extraction et transformation des données issues de sources hétérogènes afin d'alimenter des données cibles. Le transport de données devient une nécessité au sein des entreprises. Les stagiaires découvriront l'environnement de l'outil Pentaho Data Integration et apprendront à maîtriser ses fonctionnalités.

Objectifs

Comprendre l'environnement Pentaho et son fonctionnement
Maîtriser l'utilisation d'un ETL (Extract Transform Load)
Maîtriser l'outil Pentaho et la bibliothèque de composants

Programme détaillé

PRÉSENTATION

Pourquoi Pentaho Data Integration (PDI) ?

Historique et présentation au sens large

Installation et configuration de l'outil

GÉNÉRATION D'UNE PREMIÈRE EXTRACTION DE DONNÉES

Les transformations et tâches (ordonnancement des transformations)

De l'extraction à l'alimentation de données

Appréhender et gérer les flux de data

Exécution d'une transformation d'une tâche

ACCÉDER AUX DONNÉES SOURCES / CIBLES

La notion de métadonnées

Configurer l'accès aux sources de données

Les sources / cibles supportées

Les liens entre les sources (jointures)

Alimentation en Insert / Update (Insertion / Mise à jour)

MANIPULER LES DONNÉES

Trier son flux en ascendant ou descendant

Dédoubler son flux

Filtrer ses données selon plusieurs critères (alléger le flux)

Extraction d'informations à partir d'un champ (chaîne de caractères)

Remplacer une donnée par une autre

Gestion des opérateurs / opérations de calculs sur le flux

Bien utiliser le produit cartésien

Joindre les informations issues de données hétérogènes

Comparer des flux de données

ENRICHIR SON FLUX DE DONNÉES

Génération de logs

Création / Récupération des variables (dates, numériques, alphanumériques)

Utilisation du résultat d'un flux

Les propriétés d'un flux et de son ordonnanceur

LES BOUCLES

Les problématiques

Les boucles avec paramètres et le composant « Copie lignes vers résultat »

Les boucles avec les composants « Copie lignes vers résultat » et « Récupération lignes depuis le résultat »

EXPLOITATION

Gérer les erreurs

Générer des traces (logs)

Comprendre les erreurs et lancer des alertes (Débogage)

La parallélisation (exécution simultanée de plusieurs flux de données)

Import / export des développements

Automatisation des tâches / transformations

Documentation (mise en place des normes en cas d'erreurs ou de reprises)
